3. METHOD

The empirical investigation is part of a larger project entitled "Intra-Person Dynamics", which was carried out at the Center for Lifespan Psychology, Max Planck Institute for Human Development in Berlin, Germany (Li, Lindenberger, & Smith, 2005). The first central aim of the Intra-Person Dynamics Study is to examine adult age differences in patterns of intraindividual variability in four key domains of psychological functioning: Subjective well-being (i.e., affect, mood), self-regulation (i.e., motivation, subjective performance appraisals), cognition (i.e., working memory, processing speed, vigilance), and sensorimotor performance (i.e., postural control). The second aim of the project is to examine the dynamical day-to-day coupling between these four domains as an extension of prior research on cross-domain associations, which has largely been restricted to cross-sectional and long-term longitudinal data.

3.1 General Procedure

3.1.1 Intra-Person Dynamics Study Design

The general study design consisted of three main parts: a baseline assessment, a daily assessment phase spanning 45 sessions, and a posttest assessment (Figure 3.1). All testing was carried out at the Max Planck Institute for Human Development. Each session was conducted in a standardized manner by one of two doctoral students in the project or a trained research assistant. For practical reasons given the staffing of intensive daily data collection and availability of two laboratories with a balance platform, data collection spanned about one year, with three "waves" of daily data collection. Each wave included up to 20 participants (see Section 3.3). The first wave took place between October and December 2003, the second wave between January and April 2004 and a third group of individuals participated between June and August 2004.

The project was designed to provide data for several subprojects, e.g., sensory functioning–balance, balance–working memory, emotional well-being–cognition, motivation–cognition, as well as a range of smaller microgenetic studies on single variables. The aim was to recruit equal numbers of young (20–30 years) and older adults (70–80 years) who would come to the institute on a regular daily basis for 11 weeks total (about one week baseline, 45 daily sessions across nine to ten weeks, one day posttest assessment). In addition, this study is viewed as the first exploratory investigation in a program of research that will compare dynamic models of psychological functioning in young and older adults at baseline and under testing-the-limits conditions. Furthermore, an additional group of young and older adults was recruited as a control

group, participating only in the baseline assessment and the posttest assessment, which were separated by a period of 10 weeks in analogue as the testing schedule for the daily group.

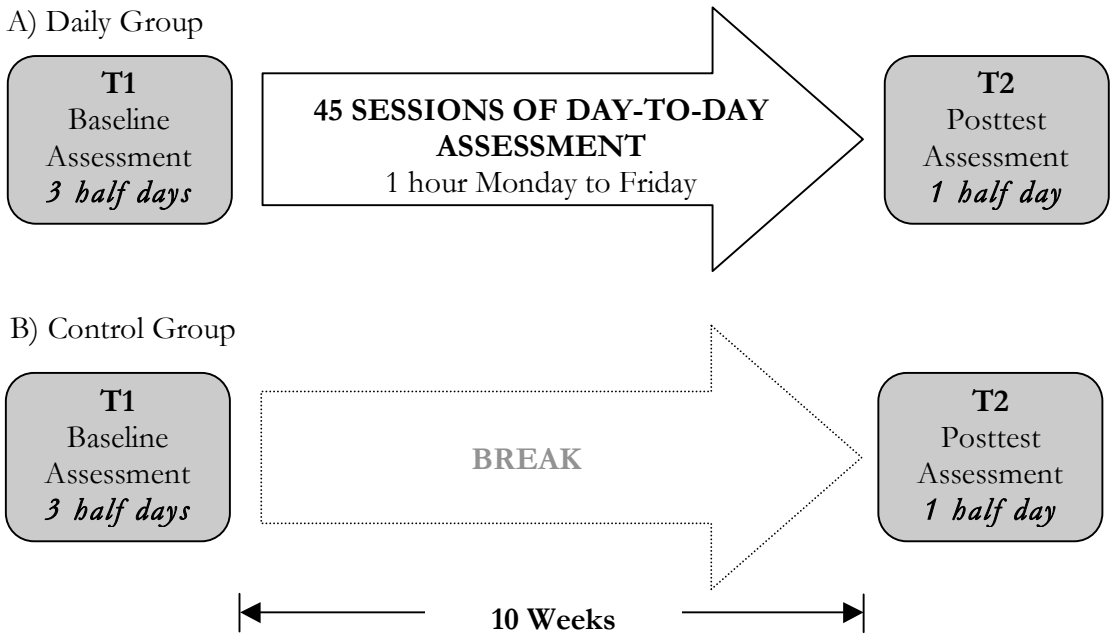

Figure 3.1. Overview of the General Study Design of Intraindividual Dynamics Project

3.1.2 Baseline Assessment

Baseline assessment consisted both of individual and small group testing (four individuals maximum), depending on the task (i.e., questionnaires versus computerized cognitive tasks) and was spread over three consecutive days. The baseline battery was intended for several purposes: sample description, possible indicators of individual differences, and potential covariates. These were three batteries: (1) self, personality, and well-being, (2) cognition, and (3) physical and sensorimotor functioning. The *baseline battery for self, personality, well-being, and health* included self-report measures of socio-economic status, well-being (affect, life satisfaction, depression, optimism, affect intensity, emotional self-regulation), self and personality (Big Five, achievement orientation, ego resiliency, control beliefs), subjective health and activities (physical self-efficacy, subjective balance, visual and auditory acuity, medical diagnoses, activities), and sleep (quantity, quality, morningness-eveningness). The *cognitive baseline battery* comprised psychometric measures of intellectual functioning (speed, verbal knowledge, spatial and arithmetic abilities) and experimental measures of verbal and spatial working memory (two-back verbal and spatial tasks). *The physical and sensorimotor baseline battery* included standard tests of physical performance

(360° turn, 3m-walking, blood pressure, vision, hearing) as well as experimental tests of postural control measured on a force platform.

In the very first testing session at baseline, participants were initially given a short presentation on a computer screen explaining the overall study design and major components of the baseline and the daily assessments. Following this introduction, participants signed an informed consent form in which they declared that they were willing to participate in the study, including 45 daily sessions, and that they had been informed about their right to resign from participation at any time without further explanation.

Participants then completed a first set of questionnaires concerning well-being, self, and personality measures. For reasons concerning the availability of testing rooms, and to accommodate for parallel testing, the entire baseline battery was split into different modules (e.g., questionnaires, cognitive testing, postural control testing, physiological health assessment) that were combined into four different orders (spanning three half days of testing: four hours, four hours, and three hours on the first, second, and third day, respectively), one for each participant tested within the same half-day baseline assessment slot. Participants were assigned to one of the four tracks depending on how long they took to fill out the initial questionnaires. For example, the person to finish first was assigned to the track that would potentially last the longest in order to ensure that no participant would be tested over-time.[8] Participants received the total amount of €110 (i.e., €10 per hour) for the baseline assessment at the end of the third day of testing.

3.1.3 Daily Assessment

In between baseline and posttest sessions, participants took part in a daily testing phase. Table 3.1 gives an overview of the overall daily assessment protocol, thereby highlighting the central measures (i.e., affect and cognitive performance) of this study in bold and placing them in the context of the overall daily session.

The one-hour daily testing sessions took place between 9 a.m. and 8 p.m. Time of day of the daily test sessions was self-selected by the participants according to their preferred timing and kept constant across the nine weeks to the greatest possible extent. This was supposed to control for age-related differences in optimal testing time and general time of day effects possibly interacting with performance particularly in the cognitive tasks (May, Hasher, & Stoltzfus, 1993; R. West, Murphy, Armilio, Craik, & Stuss, 2002b). In general, young participants were quite

[8] Prior pilot testing had informed planning of the baseline testing of the main study in terms of how much time younger and older adults would require for each module.

evenly distributed across the hourly time slots, whereas older adults were mainly tested during the morning and early afternoon (see Table B1 in Appendix B for the exact distribution of participants across time slots). Consistent with this, Levene's test of equality of error variances was significant (F (1, 35) = 12.27, p < .001). However, a univariate analysis of variance indicated that the average time of testing did not differ between the age groups (F (1, 35) = 1.44, p = .24).

Table 3.1
Daily Assessment Protocol (Sessions lasted 1 hour)

No.	Construct/Task	Approximate Time Needed
1a	**Emotional Well-Being: Positive and Negative Affect**	5 minutes
1b	Daily Stress, Daily Events, Sleep Quality	2 minutes
2	Blood Pressure	2 minutes
3	*Pretask Performance Appraisal & Motivation: Vigilance*	1 minute
	Vigilance Task	5 minutes
	Posttask Performance Appraisal & Satisfaction: Vigilance	1 minute
4	*Pretask Performance Appraisal & Motivation: Working Memory*	1 minute
	Working Memory Task	15 minutes
	Posttask Performance Appraisal & Satisfaction: Working Memory	1 minute
5	*Pretask Performance Appraisal & Motivation: Balance*	1 minute
	Balance Task: Single Task Condition	12 minutes
	Posttask Performance Appraisal & Satisfaction: Balance	1 minute
6	Digit-Symbol Substitution Test	1.5 minutes
7	Balance Task: Dual Task Cognition (with WM Task)	12 minutes

Note. Variables/tasks central to dissertation highlighted in bold.

The literature on time of day effects interacting with age in predicting cognitive performance suggests that older adults' optimal testing time is in the morning and early afternoon and that they are more severely affected by being tested outside of their optimal testing time than young adults (May et al., 1993; R. West et al., 2002b). Given this, we consider our procedure of

self-selected time to have yielded a satisfactory distribution of age groups across time slots[9]. In order to accommodate for sessions missed during the first nine weeks of the study, an extra week was added during which participants who had missed a previous session came for the respective number of make-up sessions (i.e., up to a total of five make-up sessions). The one-hour daily laboratory sessions consisted of short measures corresponding to each of the four key domains of psychological functioning outlined above.

The sequence of questionnaires and tasks was invariant across the 45 sessions and across participants. Short versions of the general task instructions were repeated to participants in each daily session, emphasizing the main important points.[10] Participants received individual daily feedback on three out of four daily cognitive/sensorimotor tasks. The feedback consisted of providing participants with their mean reaction time on correct responses and their overall accuracy (vigilance task, working memory task) and of their mean center of pressure area (balance task). Thus, the feedback provided participants with the opportunity for within-person but not between-person comparisons.

3.1.4 Posttest Assessment

At posttest, which took place 10 weeks after the last baseline session, a shorter version of the baseline battery was reassessed. The purpose of this posttest assessment was to examine whether repeated assessment had any effect on the trait measures of well-being and particularly whether any learning could be observed in cognitive performance measures which were highly similar to the ones assessed daily (i.e., close-transfer) and those that measured other constructs (i.e., far-transfer). Only some data from this posttest assessment were relevant to the present dissertation study, including information on how typical the daily assessment period was for participants' lives in general and trait-like re-assessments of emotional well-being. At the very end of the study, participants received €50 for the posttest session.

[9] Testing time was uncorrelated to a measure of diurnal preference (i.e., morningness-eveningness). There was a trend for a positive relationship between eveningness and testing time (r = .30, p = .07), suggesting that individuals with higher values on self-reported eveningness were also tested during later times of the day. Given that there were significant age group differences in eveningness indicating that younger adults scored higher than older adults on this measure (F (1, 35) = 15.83, p < .0001), the correlation between testing time and eveningness suggests, that the slight age-related differences in distribution of testing times were consistent with age-related differences in the propensity for eveningness.

[10] Primarily members of the younger subsample very quickly got annoyed with this repeated giving of the instructions. Research assistants had been trained to respond to this when it first occurred by emphasizing the importance of a standardized procedure during testing, and that because some participants would still require instructions, it was given to all participants, even if they thought they did not need it anymore. Generally, all participants appeared satisfied with this explanation and the procedure during the remainder of the study.

3.1.5 Recruitment

The task of finding volunteers interested and able to participate over 45 days for 1-hour a day is not easy. For this reason, participants were recruited using multiple strategies: through newspaper and radio advertisements, from the participant pool at the Center for Lifespan Psychology, MPI for Human Development, and through referrals of other participants already recruited for the study. During recruitment, potential participants were told that the study's aim was to examine everyday cognitive and physical functioning in different age groups and would consist of an intensive data collection period that would require them to come to the lab each weekday for nine weeks. Furthermore, they were informed about the monetary compensation that they would receive for each testing hour and as an additional bonus at the end of the data collection period. Because initial inspection of the data suggested that several participants reported relatively high depression scores, beginning with the second wave of data collection, interested volunteers were screened for a diagnosis of depressivity or depression by phone. Interested individuals with such a diagnosis were excluded from participation.

3.1.6 Strategies Used to Minimize Attrition

Several strategies were used to minimize attrition across the 45 days. First, participants received €10 per hour paid each Friday. During the daily assessment part of the study, participants were paid on a weekly basis, thus receiving up to €50 each Friday. They were also told that they would receive an additional monetary bonus of €200 under the condition that they would at least provide 40 days of data[11]. Second, participants signed an informed consent form that included an emphasis on their willingness to participate for 45 daily sessions. In this regard, the informed consent form was treated like an informal contract between participants and experimenters. A third strategy was to emphasize the personal contact between participants and experimenters. To this end, scheduling of experimenters to test sessions attempted to maximize continuity and consistency in participant-experimenter dyads. In addition, the principal investigators, including the doctoral and postdoctoral students sought regular contact with

[11] Effectively, all daily participants received this bonus, including two individuals who provided only 29 and 39 days of data, respectively. The young woman who dropped out of the study after 29 days only did so because she found a job during the course of our study, which no longer allowed her to come to our lab. Initially she made an effort to continue coming early in the morning, at 8 a.m. instead of 9 a. m., but eventually she decided that this schedule was still too stressful for her. She received the bonus because she had wanted to participate and had made an effort to do so. The older man who provided only 39 days of data missed the initial five sessions due to sickness and the missed an additional three sessions throughout the course of the study due to a vacation that he had already announced when he was recruited. Together with the available make-up days during the first wave of data collection to which he belonged, the maximum number of sessions he could possibly provide was 39.

participants throughout the course of 45 sessions. Fourth, there were several motivational events such as homemade muffins and cake each Friday (in addition to the regular cookies, tea and juice provided on each session), a card and a gift certificate for a bookstore at each participant's birthday, as well as small seasonal gifts (chocolate Santa Clause/Easter cookies). Fifth, upon scheduling the respective daily time-slots for each volunteer, participants were given the possibility to choose their preferred time of day and scheduling attempted to acknowledge these preferences in the best possible way. As such, participants were asked to try to integrate the daily testing sessions into their overall daily schedule, much like a work-like routine.

These strategies seemed to work quite well as indicated by the very high participation rates (see Table 3.2). Of all possible 1665 (37 participants × 45 days) daily measurement occasions, 1649 (i.e., 99.04%) were obtained[12].

Table 3.2
Participation Rates in Daily Sample (N = 37)

n Sessions	*n* Participants	%
39	1	2.7
42	1	2.7
43	2	5.4
44	3	8.1
45	30	81.1

3.2 Specific Design, Procedure and Measures for the Dissertation Study

Figure 3.2 illustrates the specific design and central constructs for this study. Only the main constructs assessed at each occasion are listed below the design for illustration. Within the daily assessment phase, participants first responded to a questionnaire that included the central daily emotional well-being measures as well as items on self-rated stress and daily events. Although the general procedure and ordering of questionnaires and tasks remained invariant across sessions (see Table 3.1), three versions of the well-being sub-questionnaire were created with different random orders of the items. This was done to counter-act order-effects and a

[12] Two young participants even provided 46 data points on some but not all measures. In one case, the participant reported having a headache after the session, and deliberately returned for a 46th session at the end. In the other case, the power went off during the course of a session, leaving that session with available data on only some dimensions, and therefore the participant was asked to come back for an additional 46th session. Because the opportunity for 46 occasions was an exception and not generally provided to participants, the number of sessions were set to be 45 for these two young participants for the purpose of computing the participation rate reported in the text.

decrease in range of intraindividual variability due to extensive familiarity with item order (e.g., Shifren et al., 1997). The sequence of questionnaire versions across the 45 days was varied randomly, but did not differ across participants. In the next sections, those measures specifically focused on in the present dissertation will be described in more detail (Tables B3 and B4 in Appendix B provide an overview of the descriptives of the central and control/background measures, including mean, standard deviation, minimum, maximum, skewness, and kurtosis).

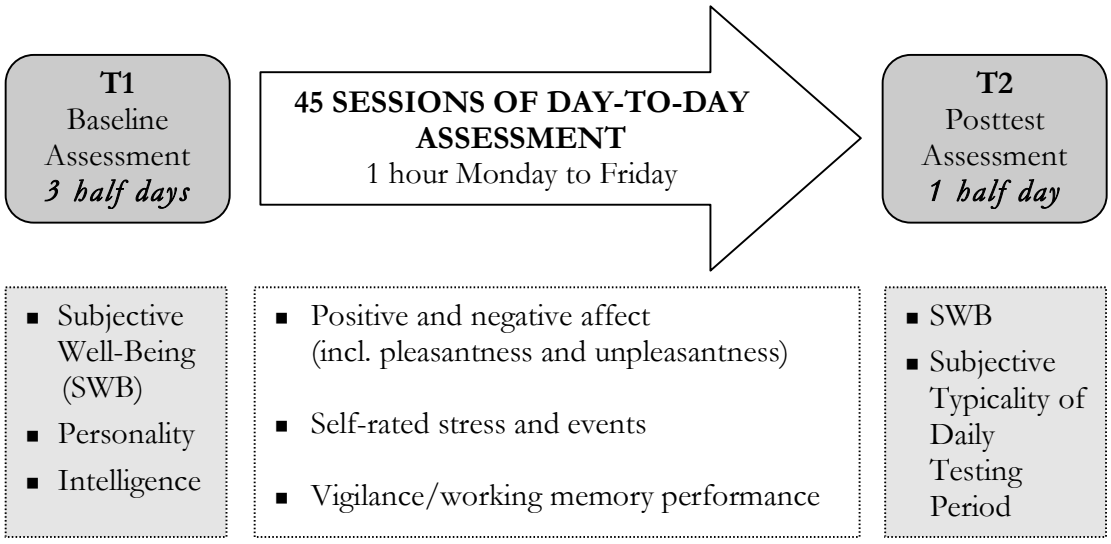

Figure 3.2. Overview of Specific Design and Central Constructs for the Present Study

3.2.1 Baseline Assessment of Central Trait-Like Covariates

The trait covariates were assessed during the baseline testing sessions. They are presented in the order of appearance in the Results chapter: First, personality factors as individual difference predictors of affect fluctuation over and above chronological age, and second, indicators of trait well-being and positive psychological functioning used to examine the functionality of affect fluctuation (see Table 3.3).

Personality: Extraversion and Neuroticism

Big Five personality factors were measured with the German version of the NEO-Five Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 1993). The questionnaire involves a total of 60 items that assess the five factors neuroticism (N), extraversion (E), openness (O), agreeableness (A), and conscientiousness (C). For the present dissertation, only the subscales assessing E and N were used. Considering the degree to which each statement described them,

participants responded to each item on a 7-point scale ranging from 1 (does not apply at all) to 7 (applies very well). Items were averaged within subscales (after recoding of appropriate items) to obtain subscale averages per participant. Internal consistency for the E and N subscales were Cronbach's α = .78 (*M* = 4.11, *SD* = 0.44) and α = .87 (*M* = 3.54, *SD* = 0.51), respectively.

Trait Subjective Well-Being and Positive Psychological Functioning

Trait or habitual positive and negative affect was measured with the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), using trait-instructions. Additional positive (happy, content, cheerful) and negative (sad, downhearted, frustrated) affect items from the emotion literature (e.g., Carstensen et al., 2000; Diener & Larsen, 1984; Watson & Clark, 1994) were included in the questionnaire in order to cover individuals' habitual experience of broader pleasant and unpleasant emotional experiences (corresponding to the daily instrument, see below for a more detailed outline).

In the trait-version of the questionnaire, a response scale ranging from 1 (never) to 7 (very often) was used to rate the frequency of experiencing each emotion during the past year. The PANAS positive affect scale (*M* = 5.17, *SD* = 0.57) had an internal consistency of Cronbach's α = .77. The PANAS negative affect scale (*M* = 3.13, *SD* = 0.92) had an internal consistency of Cronbach's α = .84. Internal consistency for the pleasantness subscale (*M* = 5.22, *SD* = 1.02) was Cronbach's α = .86, and for the unpleasantness subscale (*M* = 3.15, *SD* = 1.23), it was α = .82.

The cognitive component of subjective well-being, life satisfaction was assessed using the Satisfaction With Life Scale (SWLS; Pavot & Diener, 1993), a 5-item measure involving statements on satisfaction with one's life and living conditions. Responses were given on a 7-point scale ranging from 1 (does not apply at all) to 7 (applies very well), indicating the extent to which each item described the participant in general. Averaging across items yielded the mean life satisfaction score for each participant (*M* = 4.65, *SD* = 0.96). Internal consistency for the scale was Cronbach's α = .78.

Positive psychological functioning as a broader aspect of subjective well-being was measured using the 54-item version of a questionnaire developed by Ryff (1989; German translation by Staudinger, Fleeson, & Baltes, 1999). The instrument consists of six 9-item subscales that measure the degree of (a) environmental mastery (i.e., having a sense of being able to master one's life and the environment), (b) autonomy (i.e., living in a self-determined and independent way, having one's own standards), (c) personal growth (i.e., believing in meaningfulness of own life, having goals in life), and (f) self-acceptance (i.e., having positive

attitude towards self, accepting positive and negative aspects of self). Considering the extent to which each item described them in general, participants responded to each item on a 7-point scale ranging from 1 (does not apply at all) to 7 (applies very well). Items were averaged (after recoding appropriate items) within each subscale to obtain mean scores for each facet of positive psychological functioning. Internal consistencies were Cronbach's α = .88 for the Environmental Mastery subscale (M = 5.16, SD = 1.04), α = 0.73 for the Autonomy subscale (M = 4.91, SD = 0.76), α = .74 for the Personal Growth subscale (M = 5.57, SD = 0.71), α = .73 for the Positive Relations subscale (M = 5.49, SD = 0.81), α = .72 for the Purpose in Life subscale (M = 5.41, SD = 0.82), and α = .88 for the Self-Acceptance subscale (M = 5.31, SD = 0.98).

Aggregation of Subjective Well-Being Measures

The different measures of well-being were aggregated consistent with the well-being literature for reasons of parsimony (e.g., McGregor & Little, 1998; Ryan & Deci, 2001; Ryff & Keyes, 1995): Pleasantness, unpleasantness, and life satisfaction were aggregated into an index of contentment/happiness, the six subscales of the Ryff Questionnaire were combined into an index of positive psychological adjustment.[13]

Prior to aggregation, mean unpleasantness scores were subtracted from mean pleasantness scores to form a hedonic tone/balance score (possible range: –7 to +7, with zero indicating perfect hedonic balance, and positive scores indicating that the overall hedonic tone is positive; M = 2.06, SD = 1.95). The two subscales were negatively correlated (r = –.51, p < .01), and the resulting hedonic tone index was also significantly correlated with life satisfaction (r = .61, p < .01). Internal consistency of contentment/happiness was satisfactory with Cronbach's α = .66.

Consistently, intercorrelations between the six subscales from the Ryff Questionnaire were moderate to high, ranging from r = .35 (p < .05) between Autonomy and Self-Acceptance to r = .81 (p < .01) between Environmental Mastery and Self-Acceptance. Exceptions were the associations between the Autonomy scale and Personal Growth, Positive Relations, as well as Personal Growth (all ps > .10). The lack of strong relationships involving the Autonomy subscale may have been due to the small sample size, therefore subscale scores were aggregated nonetheless for reasons of parsimony. Internal consistency of positive psychological adjustment was α = .84 (M = 5.31, SD = 0.64).

[13] The PANAS PA and NA subscale scores were excluded from aggregation to prevent item overlap in analyses, which examined associations between happiness and fluctuation in PA and NA based on the PANAS scale.

Table 3.3
Overview of Central Baseline and Daily as Well as Control Instruments

Construct	Instrument	*n* Items	Authors / Source
A) Baseline: Trait-Like Covariates of Affective Fluctuation and of Affect-Cognition Coupling			
Personality			
Extraversion & Neuroticism	NEO-FFI: E & N subscales	24	Costa & McCrae (1992) Borkenau & Ostendorf (1993)
B) Daily Assessment			
Central Variables			
Emotional Well-Being			
Positive and Negative affect	Positive and Negative Affect Schedule (PANAS)	2 × 10	Watson et al. (1988)
Pleasantness and Unpleasantness (Hedonic Balance)	Additional items from PANAS-X and from emotion literature	In total: 6	Watson & Clark (1994), Carstensen et al. (2000), Diener & Larsen (1984)
Cognitive Performance			
Vigilance (mean RT hits)	Auditory Oddball Task	200	cf. Näätänen (1990)
Working Memory (mean RT hits)	Spatial 2-back task with additional processing demands	4 × 22	Kwon et al. (2002; modified)
Covariates			
Daily Event Covariates			
Stress	Perceived Daily Stress	1	Project
Positive and negative events	Subjective Daily Positive & Daily Negative Events	2	Project

(Table continues)

Table 3.3 *(continued)*

Construct	Instrument	*n* Items	Authors / Source
C) Baseline/Posttest: Additional Sample Descriptive and Control Variables			
Subjective Well-Being			
Positive & Negative Affect	Positive and Negative Affect Schedule (PANAS)	20	Watson et al. (1988)
Global Life Satisfaction	Satisfaction With Life Scale (SWLS)	5	Pavot & Diener (1993)
Depression	Center for Epidemiological Studies Depression Scale (CES-D)	20	Radloff (1977); Hautzinger (1988)
Positive Psychological Functioning	Ryff Inventory	54	Ryff (1989)
Emotional Experience			
Affect Intensity	Affect Intensity Measure (AIM)	40	Larsen & Diener (1987)
Intelligence			
Perceptual Speed	Identical Pictures	–	Lindenberger et al. (1993)
	Digit-Symbol Substitution Test (WAIS-R)	–	Wechsler (1981)
Verbal Knowledge	Spot-a-Word	–	Lindenberger et al. (1993)
	Vocabulary (WAIS-R)	–	Wechsler (1981)
Motivation and Experience During Participation			
Achievement Motivation	Personality Research Form (PRF): Achievement Motivation Subscale	32	Stumpf et al. (1985)
Self-Rated Representativeness of Daily Testing Period	Typicality of Nine Weeks Item	1	Project
Sociodemographic Characteristics	Age, sex, marital status, education, current occupation	5	Project

3.2.2 Central Daily Measures

Positive and Negative Affect

Two central facets of emotional well-being (e.g., Diener, 1984), positive/pleasant affect (PA) and negative/unpleasant affect (NA), were measured on a daily basis. The daily questionnaire consisted of a total of 46 emotion adjectives. This measure was designed to take no longer than 3–5 minutes. Participants were instructed to respond to all items by indicating the extent to which they experienced each emotion "right now" on a scale ranging from 1 (not at all) to 8 (very strongly).

For reasons of parsimony, 26 emotion items were selected for analyses in the present dissertation. Twenty adjectives were a German translation of the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), with ten positive and ten negative affect items. Positive affect items are active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, and strong. Negative affect items are afraid, ashamed, distressed, guilty, hostile, irritable, jittery, nervous, scared, and upset (see Table B2, Appendix B, for the English-German translation of each item). The PANAS was selected because it has high levels of validity and reliability (e.g., Watson et al., 1988), and numerous studies (most of which were conducted with younger adults) have shown this instrument to be sensitive to short-term fluctuations. Despite the wide usage of Watson and colleagues' conception of PA and NA and of their instrument itself, several criticisms of the PANAS have been voiced (e.g., Larsen & Diener, 1992; Larsen & Kasimatis, 1991). The two dimensions of positive and negative affect that are proposed by David Watson and colleagues and that are included in the PANAS represent positive and negative activation (Watson et al., 1999) more so than a pure dimension of pleasantness and unpleasantness. For example, items such as happy or sad are not included in the scale, even though most lay people tend to see these as prototypical to subjective experiences of positive and negative emotions (cf. Larsen & Diener, 1992). The PANAS was constructed in a way to represent two independent dimensions of positive and negative activation, rather than capturing all aspects of the emotional circumplex.

To meet this criticism, six additional emotion items from the daily assessment questionnaire were used in the present dissertation. Three of those, *happy, content,* and *cheerful*, were selected to represent pleasantness in addition to positive activation captured by the PANAS PA-subscale. The other three, *sad, downhearted,* and *frustrated*, were intended to capture unpleasantness in addition to negative activation captured by the PANAS NA-subscale. These items were selected from affect measures used in the literature, such as the extended version of the PANAS-X (Watson & Clark, 1994), the Carstensen Emotion Questionnaire (CEQ,

Carstensen et al., 2000), as well as work from Diener, Larsen, and colleagues (e.g., Diener & Larsen, 1984; Larsen & Kasimatis, 1991).

Ratings for positive and negative items belonging to the respective subscale of the PANAS were averaged on a daily basis to obtain a daily PA score and a daily NA score for each individual. In addition, ratings on the three pleasantness items were averaged to yield a pleasantness score and the three unpleasantness items were averaged to yield an unpleasantness score for each day and each participant. These two subscales capture the happiness/unhappiness component of emotional well-being not represented in the original PANAS subscales – (see Appendix B, Table B2, for English/German translations). In the Results chapter, the terms "positive affect" and "negative affect" refer to the PANAS scales and are of central interest in the present dissertation. Findings based on the pleasantness and unpleasantness scores will be clearly labeled as such and are only employed in supplementary analyses.

The average intraindividual mean for PA was M = 4.34 (SD = 1.03) and M = 1.42 (SD = 0.45) for NA. The average intraindividual mean for Pleasantness was M = 4.62 (SD = 1.12) and M = 1.55 (SD = 0.71) for Unpleasantness. Internal consistency (based on aggregated item scores across the 45 sessions for each participant) for the PANAS PA and NA subscales were α = .95 and α = .94, respectively, and α = .94 and α = .86, for the Pleasantness and Unpleasantness subscales, respectively.[14]

The PANAS scales represent positive and negative affect in terms of two independent dimensions. It is generally not expected that these scales correlate highly, especially not at an aggregated or trait-like level. Pleasantness and unpleasantness, on the other hand, are conceived as a bipolar affective continuum. Consistent with this, nine-week aggregated levels of PANAS PA and NA subscales were not reliably associated (r = –.21; *n.s.*), whereas the intercorrelation of the Pleasantness/Unpleasantness scores was significantly negative (r = –.58; $p < .0001$). Therefore, the latter were aggregated to form an indicator of hedonic tone/hedonic balance by subtracting each participant's daily unpleasantness score from their respective daily pleasantness score (i.e., possible range –7 to +7, with 0 indicating neutral hedonic tone and positive scores indicating a positive hedonic tone; M = 3.07, SD = 1.64).

[14] Considering internal consistency on individual sessions instead of on the basis of item scores aggregated across the daily testing period, yielded the following Cronbach's alpha scores for the first, 23rd, and 45th session, respectively: α_{PA} = .91, .93, .92, α_{NA} = .87, .74, .89, $\alpha_{PLEAS/HAPPINESS}$ = .89, .85, .82, and $\alpha_{UNPL/UNHAPPINESS}$ = .90, .72, .61.

Cognitive Performance

Most theoretical accounts on the relationship between affect and mood on general cognitive abilities are based on the idea that similar resources are needed for both domains, and the nature of these resources are thought to be attention and working memory. Therefore, a vigilance task and a working memory task were selected to assess daily cognitive performance.

The *vigilance* task was a modified version of an auditory oddball task commonly used in the attention and neurocognitive literature (e.g., Fjell & Walhovd, 2003; Näätänen, 1990). Performance on this task is considered to represent processing speed and attentional resource allocation. Specifically, a physiological indicator of performance on this task (i.e., the P300 component of electrophysiological event-related potentials, primarily latency) is related primarily to fluid components of intelligence and concentration in young and older adults (Fjell & Walhovd, 2003; O'Donnell, Friedman, Swearer, & Drachman, 1992; Walhovd & Fjell, 2002). Participants were presented with a sequence of low-pitch (800 Hz) and high-pitch (1200 Hz) tones for 4.5 minutes via headphones, while at the same time fixating a cross at the center of a computer screen. Low-pitch tones occurred at a less frequent rate than high-pitch tones and thus represented the deviants or "odds". Participants were instructed to press the space bar on a computer keyboard in front of them each time they heard a low-pitch tone, while at the same time counting the number of low-pitch tones. Thus, this task required them to attend to all presented stimuli in order to detect the deviant stimuli. At the end of the auditory presentation they were cued to type in whatever number they had counted (occasionally, for some older adults, the response was verbal and was typed by a research assistant). The interstimulus interval (ISI) between tones varied randomly from 1300 ms – 1599 ms.

The total number of stimuli was kept constant at $n = 200$ from day to day, but in order to accommodate for the repeated daily testing routine, the odds/non-odds ratio varied (i.e., the number of low-pitch tones randomly varied between 45 and 54 and the number of non-odds varied from 146 to 155 from each day to the next, always totaling 200 stimuli each day). This ratio, however, was consistent across participants on a given day. The mean reaction time for correct responses was computed as the central speed performance measure of attention on this task. One young participant showed an increase in reaction time (i.e., slower performance) across the 45-day assessment period. This person's data was thus excluded from the final analyses involving the vigilance data. In addition, one older adult did not learn the working memory task described next, and had to be dismissed from the final working memory data sample. In order to make analyses involving these two tasks comparable, they were conducted with a sample of $N = 35$. In this sample, the mean reaction time for hits was $M = 379.85$ ms ($SD = 59.85$). In

addition and for sample description purposes only, two accuracy scores were computed, but not used in the analyses of the present dissertation: Accuracy Hits, which was the accuracy of responses given hits, false alarms, and the true number of odds, and Accuracy Counts, a measure of absolute deviation from a perfect count (i.e., counting 100% of the true number of odds; Accuracy (Hits): M = .995, SD = .004, max. = 1.0; Accuracy (counts)[15]: M = .38, SD = .40, max. = 155[16]). These scores indicate that individuals' accuracy in the vigilance task was very high on average.

Working memory performance was assessed with two conditions of a computerized spatial N-Back task (modified from Kwon, Reiss, & Menon, 2002). One condition was a spatial N-back task (i.e., simple version), whereas the other condition required further processing (i.e., more complex version). The simple 2-back spatial condition was developed in our lab in analogue to a verbal n-back task commonly used in the literature (e.g., Kwon et al., 2002). On a computer screen, twenty-two dots appeared sequentially in one of eight varying locations in a 3×3-grid (no dots were presented in the central field). The stimulus interval was 2500 ms. Participants were holding two button-boxes in their hands and were instructed to press the button labeled "same" whenever a presented dot appeared in the same location like the one before last (i.e., two stimuli back). Similarly, they were instructed to press the button labeled "unequal" whenever the current position of the dot was not identical to the one before last. In this condition, 33.0% of presented stimuli were targets. The more complex version of the task ("2-back spatial plus processing"), required participants to 'mentally' move each presented dot clockwise one field over, and to compare each newly presented dot with the changed location that appeared one before last. Participants were generally instructed each day to be "as fast and as accurate as possible". Both conditions consisted of four trials each per day. For each day and participant, the average reaction time for correct responses and the overall accuracy of responses were recorded. Participants received feedback on their mean reaction time for correct responses and their accuracy separately for the simple and the complex version of the task across trials each day. In the present dissertation, only the data of the complex task version were used in order to have the maximum contrast in task demands in comparison to the vigilance task. In addition, data for one older man had to be excluded from the analyses because in the N-Back task, this person did not show any learning across time (N = 35: M_{RT} = 601.43, SD_{RT} = 309.78). Like in the vigilance task,

[15] Absolute deviation from a perfect count, hence this measure includes both counting too many and too few odds. A perfect count would hence be assigned of "0" for this measure of accuracy. The sample mean of .39 indicates, that on average participants only slightly miscounted.

[16] The maximum of 155 is theoretically possible when the true number of odds is 45 and someone counts all 200 stimuli as odds, resulting in ABS (45 – 200) = 155.

mean accuracy for this demanding working memory task was very high in the sample (M_{Acc} = .92, SD_{Acc} = .07).

3.2.3 Daily State-Like Covariates

Two types of daily state-like covariates that have commonly been linked to fluctuations in PA and NA are *daily positive events and negative events*, so-called daily hassles and uplifts, as well as *subjective appraisals of stress*. Common inventories to assess daily hassles and uplifts were too long to fit our daily assessment protocol, which had to be restricted to a maximum of 10 minutes for all self-report measures for the present study (including the assessment of subjective emotional well-being). Therefore we developed single items that assessed both positive and negative event occurrence: Using a yes/no format, participants responded to the questions "During the past 24 hours/Since the last session, did anything pleasant happen to you?" and "During the past 24 hours/Since the last session, did anything unpleasant happen to you?"[17] ('event occurrence'). Additionally, a single item assessed the degree to which participants experienced their day to be stressful ("How stressed or rushed do you feel today?" 'stress appraisal') using a response scale ranging from 1 (not at all) to 7 (very strongly). On average, individuals' stress score across the daily assessment period was M = 3.27 (SD = 1.01).

3.2.4 Additional Control and Sample Descriptive Measures

Affect Intensity

The degree of intensity with which individuals habitually experience emotions, affect intensity, was measured using the Affect Intensity Measure (AIM; Larsen & Diener, 1987). The AIM consists of 40 items that assess the degree to which one tends to experience and react emotionally in extreme ways in both pleasant and unpleasant situations, such as watching a sad film, mastering a difficult task, or being more than others excited about certain things. Individuals described their typical emotional experiences on a 7-point scale ranging from 1 (does not apply at all) to 7 (applies very much). Internal consistency of the scale was Cronbach's α = .92 (M = 4.04, SD = 0.70).

[17] The time-related reference frame "During the past 24 hours, …" was used for the very first session, whereas subsequent sessions referred to the time since the last session.

Depression

The German version of the Center for Epidemiological Studies Depression (CES-D) Scale (Hautzinger, 1988; Radloff, 1977) was used to screen for depressive symptoms. The questionnaire covers four different facets of depressivity, namely lack of well-being, depressed mood symptoms, somatic symptoms, and interpersonal difficulties. Participants were instructed to indicate the extent to which each item described their experiences during the past week using a 4-point scale that ranged from 0 (hardly ever/not at all [less than one day]) to 3 (most/all of the time [5–7 days]). After recoding where appropriate, all item ratings were summed to indicate a global depression score with a theoretical range of 0 to 60. The scale (M = 10.16, SD = 6.75) had an internal consistency of Cronbach's α = .87.

Achievement Orientation

The study design placed quite a high demand on participants' time and motivation. For sample description purposes we assessed individuals' general motivational tendencies in achievement contexts using the "Achievement Motivation" subscale from the Personality Research Form (PRF; Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1985). The scale consists of 16 items that assess individuals' habitual tendency to enjoy difficult tasks, to set high goals, and to be motivated to give their best in anything they do.

Because the original scale could have been conceived of as focusing exclusively on work-related contexts, the instruction was modified to accommodate for the fact that our sample included retired older adults. In our study, participants were instructed to think about activities they pursue in general, not only professional activities, and to indicate the degree to which each statement described them in general. The 7-point scale ranged from 1 (does not apply at all) to 7 (applies very well). After recoding where appropriate, items were averaged to obtain a mean score for achievement motivation per participant. The internal consistency was Cronbach's α = .83 (M = 4.76, SD = 0.80).

Intelligence

The cognitive baseline battery included paper-and-pencil as well as computerized tests assessing perceptual speed and knowledge. These two domains were selected to represent the two central components of intelligence proposed in the literature (Baltes, 1987; Cattell, 1971; Horn, 1982). Performance on the perceptual speed tasks indicated the fluid abilities of

intelligence, whereas performance on the knowledge tests indicated the crystallized abilities of intelligence.

The two psychometric tests of *perceptual speed* were Identical Pictures (adapted from Lindenberger, Mayr, & Kliegl, 1993) and the Digit-Symbol-Substitution Test (adapted from the Wechsler Adult Intelligence Scale-Revised [WAIS-R], Wechsler, 1981). In the Identical Pictures task, participants compare an exemplar picture with five other pictures presented simultaneously in a row below the exemplar picture. Using a touch-screen computer, participants are asked to indicate as fast as possible the one picture from the row that corresponds to the focus-picture. The task is timed to last 90 seconds. Prior to the 46 test trials, participants were familiarized with the touch screen and were given three practice trials. The sum of pictures correctly identified during the 90 seconds represents the performance indicator on this task (range: 0–46, M = 34.50, SD = 8.47). The Digit-Symbol-Substitution test is a paper-and-pencil measure in which individuals are required to substitute digits (1–9) corresponding to symbols as rapidly as possible within 90 s. The number of digits correctly assigned represents the performance score on this task (range: 0–100; M = 48.78, SD = 14.66).[18]

Knowledge was also measured with two tests: Spot-a-Word (adapted from Lehrl, 1977; see Lindenberger et al., 1993) and vocabulary (adapted from the WAIS-R, Wechsler, 1981). In the Spot-a-Word test, participants were required to choose the correct word from a list of five words containing one word and four non-words. Participants indicated the real word via touch screen. Prior to the 35 test-trials, individuals were given three practice trials. The number of correctly identified words is counted as a performance indicator on this task (range: 0–35, M = 30.0, SD = 5.33). In the German version of the WAIS-R vocabulary test (Wechsler, 1981), participants are asked to explain and define each of 16 words. Each explanation receives 0–2 points according to a coding manual, and points are summed across words to arrive at a performance index for each participant in this task (i.e., the theoretical range of scores is 0–32; M = 24.57, SD = 4.68).

Typicality of Daily Testing Phase for Everyday Life in General

In order to assess how representative the nine-week daily assessment period was for participants' lives in general, the posttest questionnaire included an item that asked "How typical were the past nine weeks, while you participated in our study, regarding your everyday life?" The response scale ranged from 1 (not typical at all) to 7 (very typical). In general, participants rated

[18] The sample mean for this task is based on N = 36, because one older participant did not understand the instruction at baseline.

the nine weeks during which the daily assessment part had taken place to be moderately typical for their life in general (M = 4.58, SD = 1.58).[19]

3.3 Participants

The first wave included nine young and seven older adults. The second wave included nine young and 11 older adults. One older woman completed the baseline assessment in Wave 2, but had to be dropped from the sample prior to the daily data collection due to difficulties with the German language (the participant was a non-native German). A third group of individuals consisted of two young and four older adults. Of the originally 43 recruited daily participants, a total of six (two young, four older) were excluded from the analyses for the following reasons: very incomplete data (n = 1 young woman), inability to learn the cognitive tasks (n = 3 older participants), drug consumption (n = 1 young woman), broken arm during the course of the study (n = 1 older woman).[20] The remaining sample of daily participants thus comprises a total of N = 37 individuals, 18 young (M = 25.50 years, SD = 2.73, range: 20–30 years) and 19 older adults (M = 74.36 years, SD = 2.86, range: 70–80 years).

Socio-demographic characteristics of the 37 participants are displayed in Table 3.4, separately for the young and the older adult subsamples. Both in the younger and in the older subsample there were almost equal numbers of men and women (young adults: 50% women, older adults: 47.4% women). The majority of the young adults were not married (83.3%), whereas the others reported being in a long-term relationship (16.7%). In the older adult group, most were married (57.4%), although as typical in this age group, this proportion was greater among men (90.0%) than among women (22.2%). Older women tended to be either widowed (33.3%) or divorced (33.3%).

[19] Inspection of the frequency of endorsed ratings indicated that one young adult endorsed a rating of 1 (i.e., not at all typical), and two young and two older adults each endorsed ratings of 2 (not typical). In addition, two young and three older adults endorsed ratings of 3. Inspection of the answers given to an open-ended question asking participants to comment on their typicality rating indicated that two young adults wrote that the nine weeks were rather untypical because they were currently writing their diploma thesis and in addition had worked a lot. Other participants did not comment. Most participants endorsed ratings of 4 and higher, indicating that the majority of both young and older adults considered the time during which they took part in the daily testing period was at least moderately typical for their life in general (see Table B5 in Appendix B for an overview of frequencies of response ratings in both age groups).

[20] These six participants did not differ significantly from the final sample of 37 participants on the basis of age (F (1, 41) = 0.57, p > .05), years of education (F (1, 41) = .01, p > .05), and subjective health (F (1, 41) = 2.03, p > .05). The only difference was based on gender, because all excluded participants were women (χ^2 (1) = 5.52, p < .05).

Table 3.4
Sociodemographic Characteristics of the Daily Dynamics Sample by Age Group

	Young Adults $n = 18$		Older Adults $n = 19$	
Age (in years)				
Range	20.89–29.49		69.91–79.21	
M	25.50		74.36	
SD	2.73		2.86	
Gender				
Male	9	(50.0%)	10	(52.6%)
Female	9	(50.0%)	9	(47.4%)
Marital Status				
Unmarried	15	(83.3%)	1	(5.3%)
Married	0	(0.0%)	11	(57.9%)
Long-term relationship	3	(16.7%)	0	(0.0%)
Widowed	0	(0.0%)	3	(15.8%)
Divorced	0	(0.0%)	4	(21.1%)
Education (in years)				
M	15.53		12.53	
SD	3.59		4.34	
Educational Level[21]				
Elementary School [a]	0	(0.0%)	3	(15.8%)
Secondary School (10th grade) [b]	1	(5.6%)	1	(5.3%)
High School (12th/13th grade) [c]	16	(88.9%)	7	(36.9%)
Technical College/University [d]	1	(5.6%)	7	(36.8%)
Other	0	(0.0%)	1	(5.3%)
Current Occupation				
University Student	14	(77.8%)	0	(0.0%)
Trainee	1	(5.6%)	0	(0.0%)
Unemployed	3	(16.7%)	0	(0.0%)
Retired	0	(0.0%)	19	(100.0%)

Notes. (a) German: Grundschule/Volksschule, (b) German: Mittlere Reife, (c) German: (Fach-)Abitur, (d) German: Fachhochschulstudium/Hochschulstudium

The majority of younger participants were university students (77.8%), whereas all of the older adults reported being retired. The educational status was slightly higher among the younger subsample than the older subsample. Most of the young adults held a German high school diploma or a higher educational certificate (94.5%), whereas 73.7% of the older adults had

[21] One older woman had classified herself as "other". Inspection of the data indicated that she had ten years of general schooling plus two years of technical university training and two years of apprenticeship. Therefore, her educational level was reclassified as High School (Fachabitur). Another older woman did not indicate her level of education. She had six years of elementary school (Volksschule) plus four years of high school (Gymnasium), but did not formally graduate due to World War II. After the war she attended a school for acting and received lessons from a private teacher. We reclassified her as "other" because given the number of years of formal and private schooling she received, her level of education is likely to be between Secondary School and High School.

graduated from high school or a higher educational institution. This suggests that both young and older adults had relatively high levels of education. More specifically, the average years of education including high school and university was higher in the younger (M = 15.53 years, SD = 3.59) than in the older subsample (M = 12.53 years, SD = 4.34; F (1, 35) = 5.22, $p < .05$, $\eta_p^2 = .13$). Therefore, number of years of education was included as a control variable in central analyses.

Table 3.5 gives an overview of descriptive information on young and older adults' self-rated health and their intellectual functioning. The majority of participants reported being in moderate to good physical health, with younger adults reporting slightly higher levels of subjective physical health than older adults (F (1, 35) = 13.39, $p < .001$, $\eta_p^2 = .28$). In terms of self-rated health, the sample was comparable with samples in previous studies (e.g., Heckhausen, Dixon, & Baltes, 1989; Salthouse & Ferrer-Caja, 2003).

Table 3.5
Subjective Health and Intellectual Functioning in the Daily Dynamics Sample by Age Group (N = 37)

	Young Adults n = 18		Older Adults n = 19	
	M	*SD*	*M*	*SD*
Subjective Physical Health	4.06[b]	.64	3.16	.83
Intellectual Functioning				
Processing Speed				
Digit-Symbol-Substitution[a]	59.17[c]	11.62	36.58	11.70
Identical Pictures[a]	42.71[d]	2.93	27.16	3.39
Verbal Knowledge				
Vocabulary	24.28[e]	3.82	24.84	5.46
Spot-a-Word	33.22[f]	5.33	26.95	3.12

Notes. (a) N = 36, (b) Range: 1 (very bad) – 5 (very good), (c) max. = 100, (d) max. = 46, (e) max. = 32, (f) max. = 35.

With respect to measures of fluid and crystallized intelligence (Baltes, 1987; Cattell, 1971; Horn, 1982), younger adults outperformed older adults on two perceptual speed tasks as indicators of fluid intelligence: Digit-Symbol Substitution, F (1, 35 = 34.67, $p < .0001$, $\eta_p^2 = .50$), and Identical Pictures, F (1, 35) = 214.34, $p < .0001$, $\eta_p^2 = .86$, a finding that is in line with the cognitive aging literature (e.g., Baltes & Lindenberger, 1997; Li et al., 2004; Salthouse & Ferrer-Caja, 2003). With respect to crystallized intelligence, age-related differences only emerged in one

of two tasks assessing verbal knowledge, favoring young adults: Vocabulary (adapted from the WAIS-R, Wechsler, 1981), F (1, 35) = .13, p = .72, Spot-a-Word (adapted from Lehrl, 1977; see Lindenberger et al., 1993), F (1, 35) = 19.36, p < .0001, η_p^2 = .36[22]. This pattern of findings is only somewhat in line with the literature, which indicates that young and older adults tend to perform equally in verbal knowledge tasks or older adults slightly outperform young adults (Li et al., 2004; Park, Lautenschlager, Hedden, Davidson, Smith, & Smith, 2002; Verhaegen, 2003).

In order to evaluate how the cognitive scores matched the norms, we compared the young and older adults of the Daily Dynamics sample with two subgroups extracted from a larger lifespan sample of 356 participants ranging in age from 6 to 89. This sample was randomly drawn from a larger parent sample of 1,920 individuals, whose contact information had been provided by the Berlin City Registry. Of those 356 individuals, 291 remained in the final sample after excluding those individuals who missed multiple testing occasions and those with severe health problems (see Li et al., 2004, for details). Data on all four cognitive tests used also in the Dynamics study were available from this reference sample. For comparison purposes with the Daily Dynamics Sample, individuals between the ages of 20 and 30 (n = 24) and those between the ages of 70 and 80 (n = 30) were selected from the 291 sample to obtain a young adult and an older adult comparison group.

A 2 (Age Group: young, older) × 2 (Sample: Dynamics, Comparison) multivariate analysis of variance with the four cognitive test scores as dependent variables yielded a significant multivariate main effect of Age Group (Wilk's lambda = .23, F (4, 82) = 69.01, p < .0001). The multivariate effects for Sample and for the Age Group × Sample interaction were non-significant (p > .05). Inspection of the follow-up between-subjects effects indicated significant main effects of Age Group for three out of the four cognitive test scores, namely for Digit-Symbol (F (1, 85) = 111.78, p < .0001), for Identical Pictures (F (1, 85) = 194.55, p < .0001), and for Spot-a-Word (F (1, 85) = 26.63, p < .0001). The main effect of Age Group was not significant (p = .78) for Vocabulary. Consistent with the results of the multivariate tests, no significant main effects of Sample and no significant Age Group × Sample interaction effects emerged for any of the four cognitive test scores in the follow-up between-person comparisons. This pattern of results allows for the conclusion that age-related differences found across tests did not differ between the two samples.

To more closely compare the two young adult groups and the two older adult groups with one another, a series of univariate analyses of variance on each of the four cognitive test scores and separately for each age group were conducted with Sample as the between-person

[22] The main effect for age group was slightly reduced but remained significant also after controlling for years of education as a covariate in the analysis of variance (F (1, 34) = 13.75, p < .001, η_p^2 = .29).

variable. These yielded insignificant main effects of Sample for all four scores and all group comparisons. In sum, results indicated that the young and older adults in the Dynamics sample did not differ significantly from their respective age group counterparts in the comparison sample.

3.4 General Statistical Procedures

All statistical analyses were conducted using SPSS 12.0 for Windows (SPSS Inc., 2003), the program HLM 5.05 (Hierarchical Linear Models, HLM Software, 2001) and SAS 9.1 for Windows (SAS Institute Inc., 2003). The specific software used will be indicated for all reported analyses. Depending on the particular analysis, either raw data or residuals were used. Unless otherwise noted, raw data were used for all analyses. However, particularly for those analyses conducted to examine within-person relationships between daily measured variables, residuals that were controlled for time-related trends in the data were used.

3.4.1 Treatment of Missing Values

In the present study, 0.34% missing data occurred at the level of individual items for the baseline and posttest measures used in the present dissertation study. For these baseline and posttest measures, missing data were not replaced. The reasoning behind this decision was that common methods for the substitution of missing data (e.g., mean imputation, regression imputation) may lead to biases in the data such as a reduction in variance, altered frequency distributions, as well as inflated covariance (e.g., Byrne, 2001; Wothke, 2000). In particular, given the small sample size, a method such as regression imputation (in which cases with complete data are used to compute a regression equation that subsequently is used for estimation of the missing value) was judged to yield biased estimates due to a lack of sufficient information. In general, missing values were very infrequent because the experimenter conducting a given session screened each questionnaire and asked the participant to complete missing responses if appropriate.[23] In addition, one of the advantages of the statistical procedures used to conduct central analyses on the daily measures (i.e., multilevel modeling) is that they are particularly apt to accommodate for missing data and allow parameter estimation on the basis of imbalanced data

[23] During the daily testing sessions, missing values only occurred in a few instances for responses to items that had a strong temporal binding and which were not of interest to the present dissertation (e.g., appraisals of performance *before* and *after* the task). Thus, participants were not asked to complete those responses at the end of the session.

sets (i.e., missing data and unequal spacing of measurement occasions across participants; Raudenbush & Bryk, 2002; Singer & Willett, 2003).[24]

Overall, on the daily level, 0.96% missing values occurred due to the fact that not all participants provided 45 sessions of data (see Section 3.1.6, Table 3.2). Of the provided sessions, 0.19% missings occurred at the level of single daily items in the self-report data (i.e., affect, stress, events), 3.0% data points and were coded as missing for the daily vigilance task due to factors outlined in the next section, and between 0.22% (older adults) and 1.51% (young adults) data points were coded as missing in the daily vigilance and the daily working memory task, respectively, according to procedures outlined below.

3.4.2 Variable Distributions and Treatment of Outliers

Using SPSS Explore, all baseline and posttest trait variables were checked for departures from normality and the existence of univariate outliers at the level of subscales. Inspection of the absolute values of skewness and kurtosis indicated that deviation from zero did not exceed the proposed cut-off values of 3.0 for skewness and 10.0 for kurtosis (Kline, 1998). Therefore, all analyses were conducted with untransformed variables.

Univariate Outliers

Outlier detection and treatment differed somewhat for the trait variables assessed at baseline and posttest and for the daily variables. For trait well-being and personality variables, following current conventions (e.g., Newton & Rudestam, 1999; Tabachnick & Fidell, 2001), scores were considered outliers if they were more than three standard deviations above or below the mean.[25] Using this criterion, a single marginal univariate outlier was detected in the aggregated working memory reaction time data. For this one person, the mean RT across the nine weeks was adjusted by assigning a score that was three SD-units above the mean.

---

[24] This advantage of multilevel modeling techniques needs to be regarded with caution, however, as it applies only to some cases of non-systematic missingness, such as "missing at random" (MAR), if not "missing completely at random" (MCAR). As described in the following section, a portion of outliers and missing data in the daily oddball assessment resulted from a PC failure that caused initial session data to be invalid or lost. This was a systematic phenomenon in the sense that it affected all participants assessed in the First Wave of data collection, but because the PC was replaced by a MAC after discovery of the problem, participants from all other waves of data collection were unaffected. However, participants in all three waves of data collection were recruited in comparable ways and assignment to each group was only determined by the temporal order of data collection and the recruitment schedule. Therefore, the missingness is unlikely to be related to central variables in the present study, as the problem was technical and not based on participants' unwillingness to respond.

[25] Most of the variables were used as Level 2-predictors in multilevel modeling analyses. Therefore, outlier detection was done on the basis of the full sample. However, some of the variables were also used with grouped data (e.g., in ANOVA, MANOVA). In those instances, outlier detection was conducted at the subgroup level (i.e., young versus older subsample).

For the *daily mood data*, inspection of raw data trajectories for each participant indicated no severe outliers. In some instances, particularly in the case of daily negative affect, a value appeared somewhat as an outlier with respect to the remaining time-series simply because most participants consistently reported very low values of NA, and any score above their overall remaining mean level of NA experiences hence stood out. Because unlike the cognitive tasks, the processes giving rise to such extreme values were not thought to differ for such values (i.e., no technical problem or responding under wrong instructions), and there was no evidence that they were artificial (i.e., in most instances, such more extreme ratings were corroborated with reports of a positive or negative event), all values were retained in the final data set to give an accurate impression of daily mood processes as they occurred in everyday life (see also Wilhelm, 2001).

To eliminate error variance in the *daily vigilance data*, outliers were identified and eliminated using the following procedure: First, the raw data were inspected for each individual to screen for obvious errors in the data that appeared to be due to technical problems or a respondent forgetting to press the spacebar on a given session. In these instances, the computer program recorded the RT to be "0". Second, extremely fast responses were defined as those RTs that were smaller than 150ms. Such responses were considered to reflect accidental key presses rather than the process under study. Third, an upper bound defined as the mean plus four standard deviations was defined for each RT data point in the individual trajectories using a moving average procedure with two different window sizes (i.e., five sessions and 15 sessions). This approach accounted for the learning effects observed in most individuals' data and was in line with the effort to separate short-term fluctuation from longer-term learning. No outliers were detected according to this last procedure. However, as a result of the first two screening procedures, a total of eight mean RTs (i.e., 0.5% of all available data points) were identified as outliers and deleted from the data set. It should be noted, however that an additional 41 mean RT values (i.e., 2.5% of all available data points) were set to missing for the first two to three sessions in the group of participants who were tested during the first wave of daily data collection. Missingness resulted from a computer problem that consisted of a recording of inflated reaction times (i.e., > 1000 ms) or of "0" response times.

Following the identification and removal of outliers, outlying values that were set to missing were imputed using SPSS Missing Value Analysis (MVA) in which estimates were computed separately for each individual based on the relationship of time in study, reaction time, and accuracy for hits and for counting correctly across days. This approach is rather conservative because it is likely that eliminating outliers and imputing them via a regression procedure results in a reduction of across-day within-person variability. Values at the beginning of a personal time-

series were not imputed via regression because inspection of the estimated values indicated that the reduction in variance was too high because the regression procedure (that included time in study to accommodate for learning over time) evidently worked less reliable for the beginning of each time-series than for values that were more embedded in each trajectory. Therefore these values were left to be missing.

For the *daily working memory data*, the procedure of detecting and treating missing data and outliers differed. Outlier detection and treatment was performed at the level of the Dynamics project and will only be summarized here very briefly: Unlike in the oddball data, no obvious technical problem was evident that warranted the deletion of certain reaction times. Because no clear criteria would distinguish extreme values from mere large day-to-day variation in performance, values were treated as outliers if inspection of the data suggested that participants did not properly follow task instructions (i.e., shift dot in clock-wise direction). Personal reports of participants had indicated that on some trials or days, some participants occasionally mixed up the simple version (i.e., no shifting of the dot) and the complex version (i.e., clock-wise shifting of each presented dot) of the task (regardless of instruction given to everybody every day), and in addition sometimes shifted the dots not clockwise but counter-clockwise in the complex condition. Thus, for each trial, the data were scored in three different ways: (a) according to complex task instructions with clock-wise shifting, (b) according to simple task instructions, and (c) according to complex task instructions with counter-clock-wise shifting. Scores were deleted if the accuracy increased by 25.0% if scored according to schemes b) and c). In those instances, it was assumed (and in most cases this corroborated reports from session protocols) that the participant had done the task under an instruction that differed from the one actually given and thus variability did not capture the process of interest. As a result of this procedure, between 0.22% (older adults) and 1.51% (young adults) of the values were identified as outlying and were deleted.

Multivariate Outliers

Multivariate outliers were identified at a between-person level as such prior to analyses if their Mahalanobis distance was significant at $p < .001$ (Tabachnik & Fidell, 2001). Using this criterion, no multivariate outliers were detected.

3.4.3 Alpha-Level Adjustment in Multiple Testing

In order to prevent alpha-level inflation, multivariate analyses were employed for a common set of dependent variables rather than conducting multiple univariate analyses.

Univariate follow-up analyses with adjusted alpha-level ($p < .01$) were conducted only when the multivariate effect was significant at the .05 level. In all other analyses, the alpha-level was adjusted to $p < .01$ to account for multiple testing. However, given the small sample size and therefore restricted power to find smaller effects, results significant at $p < .05$ are also reported, and results at $p < .10$ are reported as trends if they were corroborating theoretical hypotheses. In comparison with the common Bonferroni-adjustment, these criteria are less conservative in reducing the likelihood of finding an effect that does not really exist, but they nonetheless increase the likelihood of finding smaller effects, particularly given the small sample size and hence restriction in power (Cohen, 1990).

3.4.4 Analytic Strategy

Analyses were carried out at several different levels, including descriptive as well as inferential statistics. The nature of the day-to-day affect (and cognition and stress) data is hierarchical such that repeated assessments of daily affect and other daily variability are nested within individuals. In order to examine patterns of intraindividual variability and of intraindividual covariation, the use of a particular type of analytic tool is recommended: Multilevel modeling (MLM; also referred to as hierarchical linear models, multilevel random coefficient models). Because this analytic approach is still rather new, it will be described in more detail below, following a brief overview of the particular types of analyses carried out for each of the three sets of research questions.

Overview of Analyses Undertaken for Each Set of Research Questions

Within the first set of research questions, specifically to address age-related differences in intraindividual mean levels and in potential time-related trends in PA and NA across the daily assessment phase, zero-order correlations between aggregated mean levels of affect and age, and multilevel modeling analyses were conducted. In order to examine age-related differences in intraindividual variability in self-reported affect, several types of analyses were conducted. Initially, the intraclass correlation for both affect domains and separately for each age group was computed using multilevel modeling analyses to obtain descriptive information on the relative amount of within- to between-person variance in daily PA and NA for young and older adults. Second, for each participant, the intraindividual standard deviation in PA and NA as well as a score representing the absolute values of residuals after regressing daily affect on session were computed. The former score (ISD) represents a commonly used indicator of variability as fluctuation (e.g., Charles & Pasupathi, 2003; Eaton & Funder, 2001; Eid & Diener, 1999),

whereas the latter score represents an indicator of variability that is controlled for time-related trends in the time-series data. After deriving these variability scores from the daily time-series affect data, they were subjected to repeated measures analyses of variance (separately for PA and NA) to examine both within-person (PA vs. NA) as well as between-person (young vs. older) differences in variability as fluctuation.

Within the second set of research questions, separate sets of hierarchical regression analyses were conducted to examine the predictive value of age and personality for individual differences in variability in PA and NA. In addition, multilevel modeling analyses were employed to examine age-related differences in the day-to-day relationships between time-varying covariates such as self-reported stress and events and affect. The functional implications of individual differences in affect variability were examined on an exploratory basis (given the small sample) by computing the zero-order correlations between variability in PA and NA as well as indicators of trait psychological well-being.

Within the third set of research questions, on an initial descriptive level, the within-person relationships between cognitive performance and PA as well as NA were analyzed using separate zero-order correlations between daily cognition and affect for each individual. As a follow-up, multilevel modeling analyses were performed to examine individual differences in the day-to-day coupling between cognitive performance and affect.

Multilevel Modeling Analyses

I used the multilevel modeling (MLM) technique to examine all within-person hypotheses because it offers several advantages over classical ordinary least squares (OLS) techniques and other standard methods when assessing intraindividual variability and change (e.g., Kenny, Kashy, & Bolger, 1998; Kreft & de Leeuw, 1998; Nezlek, 2001; Raudenbush & Bryk, 2002; Singer & Willett, 2003):

First, MLM can estimate average level and average coupling/change (i.e., fixed effects) as well as individual differences in both parameters within the same model (i.e., random effects), whereas standard methods emphasize average trends. Thus, MLM techniques account for the fact that the occasions on which participants were sampled represent a random selection of a population of occasions (days), and the selection of days was meant to represent individuals' typical everyday life. Second, MLM can accommodate imbalanced designs in the sense that these techniques can handle both missing data (i.e., not all participants were assessed at all time points) and unequal spacing between measurement occasions across participants (i.e., some sessions are one day apart, others two or three, and this differs across participants; see Footnote 24 in Section

3.4.1 for a brief discussion of the missing-at-random assumption in multilevel modeling with respect to missingness in the present sample). Whereas standard approaches would have been restricted to the use of complete data and eliminate cases with missing data, using (full information) maximum likelihood estimation, MLM techniques allow parameter estimation to be based on the maximum of available data. Third, MLM techniques can accommodate for the dependency of repeated assessments within individuals and thus provide better parameter estimates than OLS methods. Due to the nature of our data, multilevel modeling techniques were used whenever the hypothesis to be examined focused on the single occasions of measurement (i.e., for individual growth models and for individual coupling models), whereas standard OLS methods were only used when the focus was on the aggregated daily data (i.e., the amount of variability across the entire assessment period).

The different levels of analysis in multilevel modeling. In essence, a multilevel model considers two levels of analysis: the within-person level and the between-person level. First, within-person relationships are estimated for each individual, yielding an intercept and a slope coefficient for each individual. Note that depending on the purpose of the model, the slope coefficient can represent the relationship between a time-varying variable (e.g., positive affect) and time itself, as is the case in individual growth models. In models that primarily focus on the coupling between two time-varying variables (e.g., positive affect and reaction time) the slope coefficient represents the within-person relationship between these two variables (i.e., the coupling or covariation between two substantial variables, with time being only represented by the repeated measurement occasions used to model the within-person coupling and individual differences in coupling).

Second, the pooled intercept and slope across participants become the dependent variables, and individual difference variables can be used to predict these pooled within-person coefficients. In essence, each model yields estimates of fixed effects (for intercepts and slopes) that describe the average sample trajectory or the average level and coupling. In addition, random effects for both intercept and level denote the average deviations from mean or starting level and around the rate of change or the coupling in person-level trajectories/relationships relative to the sample-level trajectory/relationship. Significant random effects indicate the presence of reliable interindividual differences in aspects of intraindividual level and change or coupling.

Sequence of models and model notation. More specifically to the present study, a sequence of change or coupling models was examined in order to find the model that best fitted the daily data (Singer & Willett, 2003). This sequence always included an *unconditional model* to be used as a baseline, against which the additional models were compared to in model fit and variance explained at the within-person (and eventually between-person) level. In the case of a sequence

of models examining time-related trends in the data, the so-called unconditional model can also be referred to as a no change model. Model notation differs across the literature. In the present work, the notation of Raudenbush & Bryk (2002) and their program HLM will be used because it is most commonly found in the literature. In general, the occasion-level (i.e., within-person, Level 1) component of the unconditional model is represented in the following equation:

$$\gamma_{ij} \text{ (Daily Affect/Performance)} = \beta_{0j} + r_{ij,} \qquad \textit{[Unconditional Within-Person/Level 1 Model],}$$

where γ_{ij} represents daily affect (PA or NA) or daily performance for a given person j on a given day i, β_{0j} represents the average level of affect (performance) across the nine weeks, and r_{ij} denotes a residual variance or error term representing the within-person variance in daily affect (performance) that remains after accounting for mean level affect (performance). This model assumes levels of affect (performance) to be unrelated to time or any other time-varying covariate and can thus be used as a baseline model for further model comparisons. The person-level (i.e., between-person or Level 2) component of the unconditional model is represented in the following equation:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad [\textit{Unconditional Between-Person/Level 2 Model for Intercept}],$$

where each person's intercept (β_{0j}) was modeled as a function of the grand-mean of affect (performance) across individuals (γ_{00}) and a random component (u_{0j}) representing between-person variance in levels of PA, NA, or performance that were unaccounted for by the specific affect (performance) grand-mean. In this model, the intra-person intercepts were assumed to vary across individuals. In other words, not every person was postulated as having the same average level of affect or performance across the daily testing period.

Then, several conditional models were estimated, in which daily affect (or daily performance) were analyzed at Level 1 as a function of time and other time-varying variables. Typically, separate models were run for PA and NA as outcome variables[26]. The Level 1 or within-person component of such a conditional model is represented in the following equation:

$$\gamma_{ij} \text{ (Daily Affect/Performance)} = \beta_{0j} + \beta_{1j} \text{ (Session/Daily Stress)} + r_{ij,}$$

[Conditional Within-Person/Level 1 Model].

[26] The only exception being an analysis that focused on the coupling between these two variables, which clearly required modeling both in the same model.

In a growth model, γ_{ij} represents daily affect (performance) for a given person j on a given day i, β_{0j} represents the level of affect at the beginning of the daily assessment period, because the session term was centered at the first session, β_{1j} represents the slope of the personal session number, and $r_{ij,}$ denotes the residual within-person variance in daily affect (performance) that remains after accounting for mean level affect (performance) and a linear change component. Other models tested included higher polynomials for session, such as a quadratic or cubic term. In a coupling model that was used to examine the within-person relationships between affect and stress or between performance and affect, γ_{ij} represents daily affect (performance) for a given person j on a given day i, β_{0j} represents the level of affect at the mean of the other time-varying variable(s) in the model, because these time-varying covariates were generally group-mean centered in order to control for individual differences in mean levels (i.e., centered at each person's mean across days; Nezlek, 2001). Furthermore, β_{1j} represents the slope of the time-varying covariate, which essentially denotes the within-person coupling coefficient for a given person, and $r_{ij,}$ denotes the residual within-person variance in daily affect (performance) that remains after accounting for mean level affect (performance) and a the time-varying covariate(s) such as stress or other mood terms.

The simplest between-person or Level 2 model in such a conditional model also proposes the intra-person intercepts to vary across individuals as a function of a grand mean (of level of affect/performance) and an error term, resulting in the same person-level random coefficient model for the intercepts as presented above:

$\beta_{0j} = \gamma_{00} + u_{0j}$ *[Unconditional Between-Person/Level 2 Model for Intercepts].*

In addition, however, the intra-person slopes (i.e., the trend/the coupling) are also thought to vary across individuals as a function of a grand mean (of the session/time-varying covariate slope) and an error term, resulting in the following person-level random coefficient model for the slopes (with more than one slope like in the present simplified example, there will be a Level 2 equation for each intra-individual slope):

$\beta_{1j} = \gamma_{10} + u_{0j}$ *[Unconditional Between-Person/Level 2 Model for Slopes].*

Intra-person intercepts and slopes were initially modeled as random effects. Only when problems of convergence occurred, they were modeled as fixed effects only, as represented in the following exemplar equation for a given intra-individual slope:

$\beta_{2j} = \gamma_{20}$ *[Unconditional Between-Person/Level 2 Model for Slopes, with a Fixed Slope coefficient].*

In the last step, age group and other individual difference factors were added step-by-step to the best-fitting within-person model to examine whether individual differences in intercept (level) and slope (trend, coupling) could be explained by these between-person variables. Consistent with the literature, these Level 2-covariates were grand-mean centered (Nezlek, 2001) to avoid collinearity. Age group was coded as "0" for young adults and as "1" for older adults. The following equations illustrate the between-person equations in which both the intercept and the slope is modeled as a function of a grand-mean, age group (with the younger group as the reference category), and an error term:

$\beta_{0j} = \gamma_{00} + \gamma_{01}$ (Age group) + u_{0j} *[Conditional Between-Person/Level 2 Model for Intercept]*,

$\beta_{1j} = \gamma_{10} + \gamma_{11}$ (Age group) + u_{0j} *[Conditional Between-Person/Level 2 Model for Slope].*

An unstructured covariance matrix was specified for the random effects in all models, because this is more appropriate than more structured matrices when the data structure is characterized by unequally spaced intervals (for a recent application, see Mroczek & Spiro, 2005).

Significance testing and model fit. Different strategies can be employed to test the significance of the fixed or random effects or to test whether a given model fits the data well: The simpler approach is to use a single parameter test. With respect to fixed effects, it examines the null hypothesis that the population value of a given parameter is 0 after controlling for all other predictors. With respect to random effects, it examines whether there is remaining residual outcome variation that could be explained by other predictors. It should be noted, however, that particularly with respect to the random effects, statisticians debate over the effectiveness of these tests and have questioned their utility, particularly with small samples and imbalanced data sets (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003). Therefore, the single parameter tests provide a quick method of evaluating the fixed and random effects associated with a given predictor, but despite their wide usage in published work they should be interpreted with caution. An alternative approach to single parameter tests advanced by statisticians is the use of fit statistics in comparing different models within a taxonomy. The advantage of this approach is its superior statistical properties, its ability to test several parameters simultaneously (including the comparisons of models using different numbers of random effects) and to limit the chance of committing a Type I error (Singer & Willett, 2003).

Three indicators of model fit are available for multilevel modeling. Their appropriateness depends on whether models are nested in one another or not (Singer & Willett, 2003). The most common one is the deviance statistic, which is the appropriate indicator of model fit when comparing nested models. It is a comparison of the log-likelihood statistics for a current model and a saturated, more general model, which fits the sample data perfectly. The deviance statistic is defined as the difference between these two log-likelihoods multiplied by –2, therefore it is often labeled –2 log likelihood (–2LL). It quantifies the degree to which the current model is worse than the best possible model. Thus, models with smaller values of deviance are regarded as representing better fit to the data. The difference in deviance statistics between two models has an asymptotical χ^2-distribution, where the degrees of freedom (*df*) are equal to the number of independent constraints imposed to the current model as opposed to the previous model. In order to account for the small person-level sample and the fact that despite their wide usage, single parameter tests for fixed and random effects have been criticized for being sensitive to sample size, comparisons of model fit based on the deviance statistic were employed and will be indicated in the presentation of results in addition to single parameter tests in the present dissertation.

As a measure of effect size, variance explained at both the occasion-level (i.e., within-person level) and at the between-person level is expressed with a pseudo-R^2 statistic (Kreft & de Leeuw, 1998; Singer & Willett, 2003). It has to be noted, however, that this statistic differs from traditional R^2 statistics from OLS regression analyses, and in some cases, pseudo-R^2 statistics cannot be computed or are hard to interpret. They will be reported for all analyses as a gross indicator of effect size, but should be interpreted with caution (Kreft & de Leeuw, 1998; Singer & Willett, 2003).